Latent Dirichlet Allocation: Stability and Applications to Studies of User-Generated Content

Sergei Koltcov
National Research University
Higher School of Economics
ul. Soyuza Pechatnikov, 27
St. Petersburg, Russia
skoltsov@hse.ru

Olessia Koltsova
National Research University
Higher School of Economics
ul. Soyuza Pechatnikov, 27
St. Petersburg, Russia
ekoltsova@hse.ru

Sergey Nikolenko
National Research University
Higher School of Economics
ul. Soyuza Pechatnikov, 27
St. Petersburg, Russia
sergey@logic.pdmi.ras.ru

ABSTRACT

Topic modeling, in particular the Latent Dirichlet Allocation (LDA) model, has recently emerged as an important tool for understanding large datasets, in particular, user-generated datasets in social studies of the Web. In this work, we investigate the instability of LDA inference, propose a new metric of similarity between topics and a criterion of vocabulary reduction. We show the limitations of the LDA approach for the purposes of qualitative analysis in social science and sketch some ways for improvement.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics – Probabilistic algorithms (including Monte Carlo); G.1.0 [**Mathematics of Computing**]: Numerical Analysis – stability (and instability); I.1.2 [**Computing Methodologies**]: Algorithms – analysis of algorithms; I.2.7 [**Artificial Intelligence**]: Natural Language Processing – text analysis

Keywords

Latent Dirichlet Allocation, topic modeling, social analysis

1. INTRODUCTION

With huge growth of online text data, it is becoming vitally important for social scientists to have reliable methods for fast automated analysis of such data. Researchers are, in particular, interested in methods able to track agendas, topics, opinions, and sentiments in user-generated content that can later be used for the purposes of political science, sociology, marketing, and other disciplines. One of the methods aimed at detecting topical structure in large text collections is a class of probabilistic models called Latent Dirichlet Allocation (LDA); these models have become the *de facto* standard in the field of topic modeling. However, comprehensive investigations of the quality of these models for qualitative studies are very scarce, and some indicators of quality, such as reproducibility of results, have hardly been researched at all. Instead, complex extensions of the algorithm are rapidly proliferating [2,4,9,18], as well as applications of topic modeling to specific datasets and applied goals, e.g., qualitative studies, without comprehensive prior testing [7].

Informally speaking, quality for social scientists means that the algorithm is able to show the topics "that are really there". In particular, a social scientist would expect that a topic modeling algorithm detects all "existing" topics, does not show any "non-existing" topics, and shows their "true" proportion. Then the researcher would conclude, say, whether the online public is currently talking more about elections than about popstars (in sociological context), or more about one brand than another (in marketing context). While it is unclear how to judge this notion of quality, stability is obviously an important sanity check: if a model gives different results each time it is run on the same data, it certainly does not draw the "true" picture of social reality.

In LDA, each document expresses multiple topics at once, each with a certain affinity. Likewise, each topic is a distribution on words. Thus, from the mathematical point of view each document is a mixture of distributions. To find the word-topic and topic-document matrices (probabilities of words appearing in topics and topics appearing in documents), one has to approximate the initial set of documents by these distributions. Two most popular approaches are based on variational approximations [1, 3] and Gibbs sampling [5] respectively. These algorithms find a local maximum of the joint likelihood function of the dataset; this is accepted as a solution for the topic modeling problem. Moreover, the LDA approach has been further developed by offering more complex model extensions with additional parameters and additional information [2, 4, 9, 18]. However, from the end user's point of view a local maximum does not necessarily represent a satisfactory solution for the topic modeling problem. In the case of LDA, there are plenty of local maxima [5], which may lead to instability in the output. Therefore, before using LDA social scientists have to understand how stable the output will be; this, in turn, calls for an instrument of comparison between different solutions that would be able to capture similarity between topics as sets of words with probabilities. One important problem is the huge "long tail" of words with low probabilities that are mostly irrelevant for qualitative analysis but may contribute to the level of similarity between topics. Therefore, we may need additional criteria for reducing these sets of words.

WebSci'14, June 23–26, 2014, Bloomington, IN, USA.
Copyright 2014 ACM 978-1-4503-2622-3/14/06 ...$15.00.
http://dx.doi.org/10.1145/2615569.2615680.

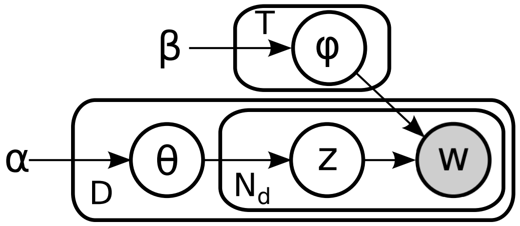

Figure 1: LDA graphical model.

In this work, we investigate the instability of the LDA algorithm, proposing a new metric of similarity between topics and a method for vocabulary reduction. We show the limitations of LDA for the purposes of qualitative analysis in social sciences and sketch some ways to improvement. Section 2 shows related work and our contributions. In Section 3, we introduce the new similarity metric, in Section 4 we use it to evaluate LDA stability, and Section 5 concludes the paper.

2. LDA AND OUR CONTRIBUTIONS

2.1 LDA

The basic latent Dirichlet allocation (LDA) model [3,5] is depicted on Fig. 1. In this model, a collection of D documents is assumed to contain T topics expressed with W different words. Each document $d \in D$ is modeled as a discrete distribution $\theta^{(d)}$ over the set of topics: $p(z_w = j) = \theta_j^{(d)}$, where z is a discrete variable that defines the topic for each word $w \in d$. Each topic, in turn, corresponds to a multinomial distribution over the words, $p(w \mid z_w = j) = \phi_w^{(j)}$. The model also introduces Dirichlet priors α for the distribution over documents (topic vectors) $\theta \sim \mathrm{Dir}(\alpha)$, and β for topical word distributions, $\phi \sim \mathrm{Dir}(\beta)$. The inference problem in LDA is to find hidden topic variables $\boldsymbol{z}$, a vector spanning all instances of all words in the dataset. There are two approaches to LDA inference: variational approximations and MCMC sampling which in this case is convenient to frame as Gibbs sampling. After easy transformations [5], Gibbs sampling reduces to the so-called *collapsed Gibbs sampling*, where z_w are iteratively resampled with distributions

$$p(z_w = t \mid \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) \propto p(z_w, t, \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) =$$
$$= \frac{n^{(d)}_{-w,t} + \alpha}{\sum_{t' \in T}\left(n^{(d)}_{-w,t'} + \alpha\right)} \frac{n^{(w)}_{-w,t} + \beta}{\sum_{w' \in W}\left(n^{(w')}_{-w,t} + \beta\right)},$$

where $n^{(d)}_{-w,t}$ is the number of times topic t occurs in document d and $n^{(w)}_{-w,t}$ is the number of times word w is generated by topic t, not counting the current value z_w.

2.2 Evaluating LDA quality with perplexity

One well established method for numerical evaluation of topic modeling results is to measure *perplexity*. Perplexity shows how well the model predicts new test samples; for a set of held-out documents D_{test} one computes $p(d \mid D) = \int p(d \mid \phi, \theta) p(\phi, \theta \mid D) \mathrm{d}\theta \mathrm{d}\phi$ for each held-out document d and then normalizes the result as $\text{perplexity}(D_{\text{test}}) = \exp\left(-\frac{\sum_{d \in D_{\text{test}}} \log p(d)}{\sum_{d \in D_{\text{test}}} N_d}\right)$. To compute $p(d \mid D)$, various algorithms have been proposed, the current standard being the so-called left-to-right algorithm [16,17].

The smaller the perplexity, the better (less uniform) is the LDA model and the more it differs from the starting distribution. However, an important drawback of evaluating the quality of a parametric LDA model with perplexity is the fact that the value of perplexity drops as the number of topics grows, so perplexity does not really yield a way to find the optimal number of topics either numerically or qualitatively. In general, topic modeling can be thought of as clustering, and it inherits certain problems of clustering, including the problem of finding the optimal number of clusters (model selection). Moreover, perplexity depends on the dictionary size which further complicates the comparison of different results. De Waal and Barnard [15] studied perplexity as a function of dictionary size (for a fixed number of topics and documents) and showed that when the dictionary was reduced by 70%, perplexity dropped by a factor of three. Unfortunately, the authors do not analyze how these changes affect the final result of topic modeling, i.e., how well the topics represent the actual contents of the dataset.

In general, perplexity is a good measure to estimate convergence of the iterative process but it is unclear how to use it to evaluate the quality of topic modeling, especially from the point of view of human interpretation.

2.3 Evaluating LDA quality with Kullback–Leibler divergence and topic correlation

Steyvers and Griffiths [6] propose to evaluate LDA quality with a symmetric Kullback–Leibler divergence. This approach is based on pairwise comparisons of two solutions to the topic modeling problem. The pairwise comparison is computed as

$$\mathrm{KL} = \frac{1}{2}\sum_w \phi_w^1 \log\frac{\phi_w^1}{\phi_w^2} + \frac{1}{2}\sum_w \phi_w^2 \log\frac{\phi_w^2}{\phi_w^1},$$

where ϕ_w^1 is the word distribution for the first topic; ϕ_w^2, for the second topic. This metric shows similarity between two topics, but further analysis that would analyze the stability of topic reproduction in multiple topic modeling experiments on the same dataset has not been performed. Besides, the Kullback–Leibler divergence only gives an estimate of the similarity of two topics while detailed analysis would have to take into account some evaluation of the *dis*similarity between two topics.

A different approach to pairwise comparisons between topics was proposed by de Waal and Barnard [15]. Instead of Kullback–Leibler divergence, they propose a method to compute correlation between documents from two topic modeling experiments. The method consists of the following steps: (1) construct a bipartite graph based on two topical solutions; (2) compute the minimal distance between topics in this bipartite graph; (3) compare topics between two cluster solutions based on the minimal distance. This means that two topics are similar if they have the smallest distance between them as compared to the distance from these two topics to other topics. To compute minimal distances in the bipartite graph, the authors use the so-called Hungarian method, also known as Kuhn's method [8]. The authors show that correlation between documents does not depend on dictionary size as much as perplexity.

2.4 Our contributions

In this work, we propose several new metrics for evaluating different aspects of topic modeling. Namely, we introduce the notions of document and word ratios that show the fraction of words and documents that are actually relevant to specific topics. This lets us drastically cut the vocabulary in our novel topic similarity metric based on Kullback–Leibler divergence; we show that this metric matches qualitative expectations of the notion of similar topics quite well. Armed with this metric, we study the stability of Gibbs sampling for LDA inference and discover that modeling results are unstable, and sociological analysis based on topic modeling should proceed with extra care. We conclude with recommendations for further studies.

In numerical experiments, we used a popular LDA inference implementation based on Gibbs sampling, GibbsLDA++ [10]. The dataset for experiments consists of Russian language LiveJournal posts for October 2013 that we have collected for the purposes of qualitative sociological and media studies. There are 298,967 posts in the dataset with 35,049,514 instances of 153,536 unique words.

3. EVALUATING SPARSITY

3.1 Word and document ratios

LDA inference algorithms based on Gibbs sampling rely upon random sampling used to generate topic variables z for document instances on each iteration. Thus, topic modeling by itself is influenced by random noise: topic variables for both documents and topics fluctuate randomly during modeling. However, the LDA inference algorithm guarantees that the iterative process converges to a certain value of perplexity with some noise, which means that the number of words and documents used in modeling also converge to a certain value.

To estimate the number of high probability words and documents, we introduce the notion of *ratio*. Ratio is closely related to the notion of perplexity. The initial distribution for words and documents is uniform, so the probability of each topic in each document starts from $1/K$, where K is the number of topics, and the probability of each word in each topic starts from $1/V$, where V is the dictionary size. During inference, probabilities of words and topics in documents change, but they still, obviously, sum up to one; some words and topics rise above the average values of $1/K$ and $1/V$, and the others sink below it.

We introduce *document ratio* as the parameter that characterizes the ratio of the total number of topics with probability greater than $1/K$ over all documents:

$$\mathrm{DR} = \frac{1}{K|D|}\sum_{d\in D}\sum_{k}\left[\theta_k^{(d)} > \frac{1}{2}\right].$$

At the beginning of the first iteration, DR = 1; over Gibbs sampling iterations, DR begins to drop and then, at some point, it stabilizes and converges to some value; we can stop the Gibbs sampling as fluctuations of DR attenuate. Similarly, we formulate the notion of *words ratio* which is the ratio of the number of words in all topics with probability higher than $1/V$ to the total number of words in all topics:

$$\mathrm{WR} = \frac{1}{KW}\sum_{w}\sum_{k}\left[\phi_w^k > \frac{1}{2}\right].$$

Note that the same document (resp., word) may participate in the computation of document ratio (resp., word ratio) several times.

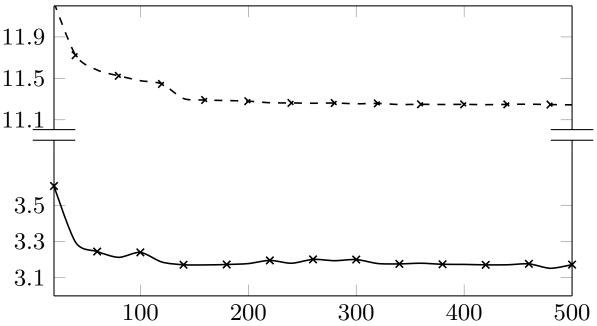

Figure 2: Sample document ratio (dashed line, %) and word ratio (solid line, %) as a function of iteration index.

Figure 2 shows the behaviour of word and document ratios for a sample run of LDA inference with 120 topics. In this case, the word ratio stabilized after 150–200 iterations around 3.2%; document ratio, around 11.5%. One can also introduce the average word ratio over a set of samples as $\mathrm{AWR} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{WR}_i$, where WR_i is the word ratio measured at the ith sample; similarly. the average document ratio is introduced as $\mathrm{ADR} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{DR}_i$. Our experiments with different number of topics (from 50 to 280) have shown that the word ratio stabilizes around 3.5% and document ratio stabilizes around 11.5% in all experiments, with standard deviation of the results being about 0.5-1%.

3.2 KL-based similarity metric

The Kullback–Leibler divergence is a widely accepted distance measure between two probability distributions. However, directly computing KL divergence to measure similarity between two topics in a topic modeling result does not lead to a good result since the KL value is dominated by the long tail of low probability words that do not define the topic in any qualitative way and are mostly random. Therefore, in this section we devise a modification for the KL metric to measure similarity between topics.

As we have shown above, the number of words with above average probabilities in our experiments was about 3.5% of the total number of unique words in all topics. We left only words top probabilities in at least one topic reducing the dictionary from 153,536 tokens (words) to 8000 (about 5.2%). This also lets us compute KL divergence faster since it has complexity $O(K^2W)$, where K is the number of topics and W is the dictionary size.

Another deficiency of the "vanilla" Kullback–Leibler divergence is that it significantly depends on the dictionary size [15]. This means that while the KL divergence is always zero (or very close to zero) when two distributions coincide almost exactly, it may have values all over the $[0, 1]$ for two very distinct topics if we consider different dictionaries and different pairs of topics, so it is hard to find a good general threshold for KL divergence. To get such a threshold, we propose to normalize KL divergence by making the distance between two least similar topics artificially equal to 1. Thus, we introduce the normalized KL similarity measure as

$$\mathrm{NKLS}(t_1, t_2) = \left(1 - \frac{\mathrm{KL}(t_1, t_2)}{\max_{t_1', t_2'}\mathrm{KL}(t_1', t_2')}\right),$$

where KL denotes the regular KL divergence. In the NLKS measure, 1 corresponds to a perfect match and 0 corresponds to the furthest possible distributions among given sets of topics.

3.3 Topic similarity thresholds

Kullback–Leibler divergence takes into account the long tail of topic-word distributions, and it may happen (and often does) that large deviations in KL-based metrics do not really correspond to significant differences in top words, i.e., the words that a qualitative researcher would use to define and understand a topic. To estimate this effect, we need to study how similarity between top words relates to the NLKS similarity measure.

Our studies have shown that in topics with similarity $0.93 - 0.95$ and higher, the 30-50 most probable words coincide almost exactly, and the sequences in which they appear in the list sorted by probability are also very similar; thus, similarity levels of 0.93 and higher indicate that a qualitative researcher would almost certainly treat these topics as the same. Similarity level about 0.9 usually corresponds to the situation when the first 30-50 words in the ranked list do match, but they have different probabilities and go in a different order; Table 1 shows a sample pair of such topics. The similarity level of 0.85 usually corresponds to a situation when two topics have a completely different set of top words.

Therefore, our experiments indicate that the proposed NLKS metric does correspond well to a qualitative estimation of topic similarity, and the similarity threshold for "truly similar" topics appears to be around 0.9. In the next section, we apply this metric to study the stability of Gibbs sampling.

4. TOPIC STABILITY

4.1 Experimental setting

In topic modeling, the posterior distribution which is maximized during inference may have a very complex and certainly nonconvex shape. This leads to multiple local maxima; in practical terms, it means that different runs of the same software may lead to different results, in particular, different word-topic distributions. Therefore, it becomes of primary importance to test the stability of topic reproduction. We propose the following method to estimate the stability of reconstructing topical solutions for given (unchanged) α and β parameters and a fixed number of topics. We perform several runs of the LDA inference software GibbsLDA++ [10] with the same parameters, getting several word-topic and topic-document distributions. Since these distributions result from the same dataset with the same vocabulary and model parameters, any differences between them are entirely due to the randomness in Gibbs sampling. This randomness affects perplexity variations, word and document ratios, and the reproducibility of the qualitative topical solution. Words may change their probabilities in topics, and it makes sense to use a KL-based measure to compare topical solutions. We use the normalized measure NLKS introduced above.

In our experiments, we performed six runs with $K = 120$ topics with model parameters $\alpha = 0.5$, $\beta = 0.1$ on our dataset with 298,967 documents and a vocabulary of 153,536 unique words. Then we performed pairwise comparisons of the results with the NLKS metric, computing how similar the topics are across different runs, for each pair of models getting a $K \times K$ matrix whose elements represent the similarity metric between topics. Then, for each topic of one model (i.e., a row of the similarity matrix) we find the most similar topic in the second model (i.e., a column).

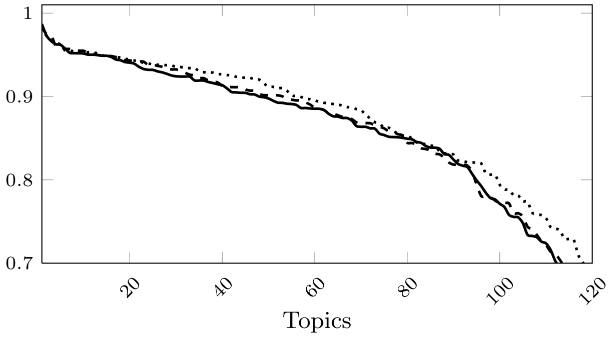

Figure 3: Topic similarity sorted in decreasing order; lines correspond to different test run comparisons.

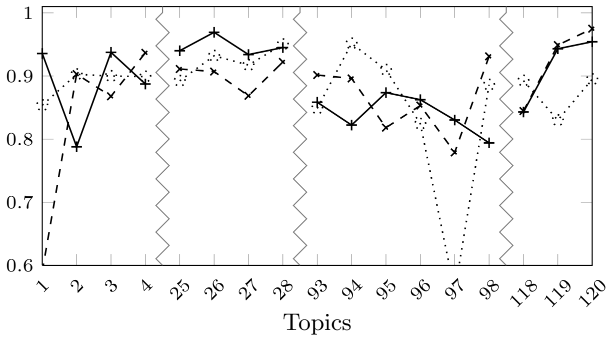

Figure 4: Sample topic similarities across test runs.

4.2 Results

Fig. 3 shows topics sorted according to similarity in three comparisons between different runs of LDA inference. It shows that less than half of the topics are reproduced with reliable stability (similarity > 0.9); this share would be even smaller if we required more than two matches. Fig. 4 shows several sample similarities between specific topics (showing the top similarity value among topics from another test). Some topics, (e.g., topics 25–28) fluctuate very little across the runs, with NLKS similarity of 0.95-1.0, while others (e.g., 1 and 97) have large deviation, with fluctuations around 40%; in practice this means that in some runs, these topics are simply not found at all. On average, fluctuations amount to 0.2065 per topic.

One might expect that the topics that do not reproduce well are "trash" topics based on common words or that would not be of interest for social studies anyway. Unfortunately, this is not the case; for instance, an interesting and readily interpretable topic on the war in Syria (first pair of topics in Table 1) reproduced only three times out of six runs in our experiments. Hence, a qualitative study might conclude that war in Syria either is very interesting for Russian bloggers or goes completely unnoticed, depending on the random number generator in Gibbs sampling.

Similarity 0.935				**Similarity 0.9**				**Similarity 0.854**			
USA	0.04734	USA	0.03567	tree	0.03195	tree	0.03321	USA	0.04734	water	0.01758
American	0.02406	American	0.01804	forest	0.021	forest	0.01918	American	0.02406	help	0.01296
Syria	0.02082	Syria	0.01758	garden	0.01527	green	0.01631	Syria	0.02082	city	0.01262
Obama	0.01374	country	0.01495	mushroom	0.015	mushroom	0.01563	Obama	0.01374	far	0.01199
weapon	0.01343	war	0.01361	leaf	0.01389	garden	0.01478	weapon	0.01343	house	0.01064
war	0.01309	military	0.01246	plant	0.01291	leaf	0.01453	war	0.01309	east	0.0104
president	0.01169	weapon	0.01084	grow	0.01146	plant	0.0135	president	0.01169	region	0.00945
UN	0.01018	Russia	0.01004	green	0.00873	grow	0.01277	UN	0.01018	dam	0.0091
military	0.01014	Obama	0.00996	collect	0.00779	color	0.01045	military	0.01014	flood	0.00904
country	0.01005	president	0.0096	rose	0.00764	flower	0.00809	country	0.01005	resident	0.00839
chemical	0.00944	UN	0.00869	flower	0.00744	rose	0.00809	chemical	0.00944	injured	0.00714
Syrian	0.00851	international	0.00769	color	0.00701	collect	0.00766	Syrian	0.00851	FRS	0.00698

Table 1: Three pairs of topics with NLKS measures. The first pair of topics did not reproduce in other runs.

5. CONCLUSION

In automated analysis of user-generated content on the Web, topic modeling provides unparallelled possibilities for sociological analysis by allowing the researcher to quickly evaluate the topical map of a corpus of texts, draw conclusions on what topics are discussed there and how intensively. However, in this work we show that classical implementations of inference in LDA models should be applied with care, since the algorithms contain inherent uncertainty in regard to which local maximum they arrive to, and unlike some other nonconvex optimization problems, in the case of LDA this does in fact matter. We show that even topics that can be easily interpreted qualitatively and appear to be full of meaning for a sociologist may be in fact unstable, showing up only in a fraction of LDA inference runs.

Therefore, to be able to draw specific sociological conclusions we recommend researchers to run topic modeling multiple times (even with the same parameters), then distinguish stable topics that reappear across multiple runs and analyze only those. We have proposed a new topic similarity measure based on Kullback–Leibler divergence.

LDA has already been critiqued for lack of stability and similar faults [11]. Our results show that further work is required to solve the underlying problem, namely to improve stability of topic modeling. One recently initiated direction of studies that we believe to be promising in this regard deals with regularized topic models. It appears that instead of Bayesian regularization it may be better to use more general Tikhonov regularizers; however, Tychonoff regularization in application to topic modeling is a research direction still in its infancy [13, 14], and further work is required.

6. REFERENCES

[1] D. M. Blei. Introduction to probabilistic topic models. *Communications of the ACM*, 2011.

[2] D. M. Blei and J. D. Lafferty. Correlated topic models. *Advances in Neural Information Processing Systems*, 18, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4–5):993–1022, 2003.

[4] A. Daud, J. Li, L. Zhou, and F. Muhammad. Knowledge discovery through directed probabilistic topic models: a survey. *Frontiers of Computer Science in China*, 4(2):280–301, 2010.

[5] T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 (Suppl. 1):5228–5335, 2004.

[6] T. Griffiths and M. Steyvers. Probabilistic topic models. In T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2006.

[7] O. Koltsova and S. Koltcov. Mapping the public agenda with topic modeling: The case of the Russian livejournal. *Policy & Internet*, 5(2):207–227, 2013.

[8] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[9] S. Z. Li. *Markov Random Field Modeling in Image Analysis.* Advances in Pattern Recognition. Springer, Berlin Heidelberg, 2009.

[10] X.-H. Phan and C.-T. Nguyen. GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA), 2007.

[11] A. Potapenko and K. Vorontsov. Robust PLSA performs better than LDA. In P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. M. Rüger, E. Agichtein, I. Segalovich, and E. Yilmaz, editors, *Advances in Information Retrieval - 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings*, volume 7814 of *Lecture Notes in Computer Science*, pages 784–787. Springer, 2013.

[12] A. N. Tikhonov and V. Y. Arsenin. *Solution of Ill-posed Problems.* Washington: Winston & Sons, 1977.

[13] K. V. Vorontsov. Additive regularization of topic models. In *Proc. 16th Russian Conf. on Mathematical Methods for Image Recognition*, page 88. MAKS Press, 2013.

[14] K. V. Vorontsov and A. A. Potapenko. Modifications of EM algorithm for probabilistic topic modeling. *Machine Learning and Data Mining*, 1(6), 2013.

[15] A. D. Waal and E. Barnard. Evaluating topic models with stability, 2008.

[16] H. M. Wallach. *Structured topic models for language.* PhD thesis, University of Cambridge, 2008.

[17] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proceedings of the 26th International Conference on Machine Learning*, pages 1105–1112, New York, NY, USA, 2009. ACM.

[18] X. Wang and A. McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433, New York, NY, USA, 2006. ACM.